REMARKS

Claims 1-3, 5-10, 14, 15, 17, 19-22, 24, 28, 30-32 and 35 are amended herein.

Claim 11-13, 25-27 are cancelled.

Claims 1-10, 14-24, and 28-37 are now pending.

Response to Rejections Under 35 U.S,C. 103

Claims 1, 2, 4, 6, 9-16, 18, 20, 23-37 are rejected under 35 U.S.C. 103(a) as being unpatentable over van den Akker (US 6,415,250) in view of Bracewell et al (US PGPub 2006/0041685). Applicants respectfully traverse.

As amended, claim 1 recites a system, comprising:

> a storage system adapted to store a set of language classes, which each identify a language and a character set encoding, and further adapted to store a plurality of training documents;
> an attribute modeler adapted to train an attribute model by evaluating occurrences of one or more document properties within the training documents and, for each language class, calculating a probability for the document properties set conditioned on the occurrence of the language class, the trained attribute model stored in the storage; and
> a text modeler adapted to train a text model by evaluating byte occurrences within the training documents and, for each language class, calculating a probability for the byte occurrences conditioned on the occurrence of the language class, the trained text model stored in the storage.

Claim 15 recites a corresponding method. These features provide a system and a method for training language identification through probabilistic analysis using two probabilistic models, an attribute model and a text model. During language identification, a set of language classes are defined, and each language class identifies a language and a character set encoding. The attribute modeler is adapted to train an attribute model by evaluating the conditional probabilistic match of select document property sets, such as a top level domain, character set Hypertext Markup Language (HTML) metatag or Hypertext Transport Protocol (HTTP) header information, and other factors, to the language classes. The text modeler is adapted to train a text model by evaluating the conditional probabilistic match of actual text to the language classes. The text is evaluated in the text modeler based on byte co-occurrences. Once the models are trained, they can be used to identify the language of documents, for example, according to the methods of claims 30 and 33.

van den Akker does not teach or suggest these features. van den Akker's language identification system identifies the language of a document based on an analysis of fixed length word suffixes. As such, van den Akker only analyzes the limited samples of text of a document, and does not have an attribute modeler as claimed.

Secondly, van den Akker does not disclose using document properties, such as character set encoding, to identify the language of a document. As claimed, a language class identifies a language and a character set encoding. For example, one language class for Japanese is "*Japanese/Shift-JIS*," where *Shift-JIS* is an encoding for Japanese Kanji characters. Some languages, such as Japanese, have multiple differing character set encodings and correspond to multiple language classes. Using the character set encoding allows the claimed system to more accurately identify a language. By contrast, van den Akker's language identification system eliminates character set encoding information entirely by converting all documents, regardless of language encoding, into ANSI text before identifying the language of the document (column 11, line 33 -39). Since van den Akker converts all languages into a single character set (ANSI), the document's original character set information is no longer available, and thus cannot be used to identify the language. As result, van den Akker does not in any way use the character set encoding as claimed. As such, van den Akker neither identifies a language class by a language and a character set encoding, nor uses an attribute modeler to calculate the conditional probability of the document properties set on the occurrence of the language class as claimed.

Thirdly, van den Akker does not analyze byte occurrence of the document to identify the language of the document. van den Akker analyzes fixed length word suffixes of words of the document. As claimed, text modeler is adapted to train a text model by evaluating byte occurrences in the documents and calculating the probability of such byte occurrences conditioned on the occurrence of the language class. For example, each trigram includes three successive bytes of word, and such trigrams support for single, multiple and variable length character set encodings, and thereby facilitates the processing of documents written in languages, such as Chinese, Japanese and Korean. By contrast, van den Akker's fixed length word suffix analysis does not disclose the byte occurrence probabilistic analysis as claimed.

Bracewell does not remedy the deficiencies of van den Akker. Bracewell discloses a system for rendering data that is not natively compatible for viewing on web browsers. Bracewell specifically teaches determining the language type from HTTP request itself (¶ 14). Thus Bracewell assumes that the language of a document is explicitly identified in the HTTP request, because only then can Bracewell select the correct template. By contrast, the claimed invention enables the identification of the language precisely in the case where the language is not explicitly identified. Thus, Bracewell does not disclose an attribute modeler that is adapted to train an attribute model to identify language attributes through a probabilistic analysis of the document properties as claimed.

The combination of teachings of van den Akker's analysis of fixed length word suffixes and Bracewell's template using language information of an HTTP request does not teach or disclose the features that train language identification through probabilistic analysis using two probabilistic models, an attribute model and a text model, as claimed. Indeed, the combination is improper under MPEP 2143.01, since it would it change van den Akker's principle of operation: Bracewell relies on the HTTP request to identify the language, and as such eliminates the need to use van den Akker's suffix analysis to identify the language of a document.

Claims 3, 5, 17 and 19 are rejected under 35 U.S.C. 103(a) as being unpatentable over van den Akker in view of Bracewell et al. applied to claims 1, 2, 4, 6, 9-16, 18, 20, 23-37 above, and in further view of Elworthy (US 6,125, 362). First, Elworthy does not remedy the deficiencies of van den Akker and Bracewell. Second, Elworthy teaches a Bayesian probabilistic formula to classify documents based on an assumed language classification and the probability of an element that is the text of the document. Specifically, in Elworthy's Equation 1, p(t) is the probability of a text token given a language classification, and thus not the probability of the document properties and actual text as claimed. Further, Elworthy does not teach that a language class is defined by a language and a character set encoding as claimed..

Claims 7 and 21 are rejected under 35 U.S.C. 103(a) as being unpatentable over van den Akker in view of Bracewell, as applied to claims 1, 2, 4-6, 9-16, 18, 20, 23-37, and in further view of de Campos (US 6,272,456). de Campos does not remedy the deficiencies of van den Akker, Bracewell and Elworthy. de Campos teaches identifying a window of letters by evaluating the frequency parameters of the matched reference letter sequences in each language profile. de Campos does not disclose or teach using an attribute modeler and a text modeler to identify language attributes through probabilistic analysis.

For at least the reasons above, Applicants submit that claims 1, 15, 30, 33, and 37 are patentable over the cited references. Claims 2-10, 16-24, 28-29, 31-32, and 34-36 either directly or indirectly depend from claims 1, 15, 30 and 33. These dependent claims also recite additional features not disclosed by the cited references. Thus, Applicants submit claims 2-10, 16-24, 28-29, 31-32, and 34-36 are patentably distinguishable over the cited references.

In sum, Applicants submit that the pending claims are patentably distinguishable over the cited references. Therefore, Applicants request reconsideration of the basis for the rejections to these claims and request allowance of them.

If the Examiner is in need of further information, he is invited to contact the undersigned attorney at the telephone number provided below.

Respectfully submitted,
Alex Franz et al.

Dated: __June 4, 2007_____ By: _____/Robert R. Sachs/___
Robert R. Sachs, Reg. No. 42,120
Attorney for Applicants
Fenwick & West LLP
Silicon Valley Center
801 California Street
Mountain View, CA 94041
Tel.: (415) 875-2410
Fax: (415) 281-1350